THE G/G/2 QUEUE:
CYCLIC vs. FIFS SERVICE ORDER

by

REINHARD PIATER
M.S.E.E., Polytechnic Institute of Brooklyn, 1965

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

in the Department
of
MATHEMATICS
and the
INSTITUTE OF APPLIED MATHEMATICS AND STATISTICS

We accept this thesis as conforming to the
required standard.

THE UNIVERSITY OF BRITISH COLUMBIA
August, 1975

Department of Mathematics

The University of British Columbia
2075 Wesbrook Place
Vancouver, Canada
V6T 1W5

Date August 31, 1975

ABSTRACT

The relative waiting times in a G/G/2 queue are investigated for FIFS vs. cyclic service order. We will prove that under the FIFS system the expected wait is shorter given rather weak conditions on the arrival process. However, the wait is not necessarily stochastically less, nor is the average wait less for every realization. This result bears on the upper limits on expected wait in a G/G/k queue given by Brumelle and Kingman.

TABLE OF CONTENTS

Page

Abstract ii

Table of Contents iii

List of Figures iv

Acknowledgement v

Text:

I. Introduction and Summary 1

II. The Two Systems - Definitions and Elementary Notions 5

III. The Two Systems - Some Partial Results 9

A. Comparison of Realizations 9

B. Comparison of Distribution Functions for Waiting Times and Total Work 11

IV. Proof of $EW \leq E\tilde{W}$ via Value Functions 15

Bibliography 25

Appendix A: Proofs of Realization Theorems 26

Appendix B: Distribution Functions for Waiting Times and Total Work 31

LIST OF FIGURES

Page

Fig. 1. Geometric Depiction of Work Loads 6

Fig. 2. Pictorial Representation of Counterexample 29

Fig. 3. Ψ -Set for (W,Q) 31

Fig. 4. Ψ -Set for (U,L) 34

Fig. 5. An Example of the Non-Monotonic Behavior of U 36

ACKNOWLEDGEMENT

This work was done under the supervison of S. L. Brumelle of the University of British Columbia, to whom the author is grateful for this interesting problem and for his patient encouragement during periods when the problem appeared to be intractable.

Reinhard Piater

I. Introduction and Summary

A familiar example of a queuing system is the servicing of customers at a bank. Customers arrive at the bank according to some probability law. There are k tellers working in parallel. A given customer either gets served immediately by one of the tellers or else he first waits in some line. The successive time spans marked off by the arrival of the customer, his beginning of service, and end of service are called the waiting time and service time. The method whereby a customer is assigned to a given teller is called the service discipline. For example, banks in Vancouver, B.C. prior to 1974 had the following service discipline: there were k separate queues, one in front of each teller, and a given customer would choose one of the queues (presumably the shortest) immediately upon his arrival. The customer could also leave the bank without being served (this is called "balking") or he could switch queues during his waiting period ("jockeying"). During 1974 some Vancouver banks switched over to the single queue system wherein the person in front of the queue would go to the next available teller. This service discipline is called "first-in-first served" (FIFS).

Although most phenomena that occur in real-world queuing systems (jockeying,balking, bulk arrivals, specialized servers, etc.) have been considered and analysed in the queuing literature, substantial attention is still being paid to the standard classical model, which may be described as follows: The service discipline is FIFS, with no balking, the interarrival times are independent, identically distributed (i.i.d.) random variables independent of the service times, which themselves are i.i.d. random variables, and the k servers all work at the same rate, independent of the length of the queue. The behavior of this system depends only on the distribution function of the service times, H(s), and that of the interarrival times, G(t).

Of particular interest are the waiting times of customers, and indeed a good figure of merit for a queuing system is the limiting expected waiting time. In general, this quantity is difficult to calculate, even for a single server queue (k=1). Marshall [1] and Kingman [3] have developed an upper bound for the expected waiting time in a single-server queue in terms of the means and variances of the service and interarrival times. For $k > 1$, the situation is considerably more complicated. However, Brumelle [2] and Kingman [3] have succeeded in establishing upper bounds for the expected waiting time in a k-server queue by comparison to a judiciously chosen single-server queue in which the expected waiting time is greater. Brumelle's single server queue was constructed from the k-server queue by assigning all arrivals in the latter to the first server, but giving those arrivals that would not have gone to the first server a zero service time. Under this modification none of the waiting times is decreased[1].

Kingman, on the other hand, asserted that the expected waiting time in a k-server queue cannot decrease if the service discipline is changed from FIFS to one where the arrivals are assigned cyclically to the k servers. (Thus the i^{th} server would get the i^{th}, $(i+k)^{th}$, $(i+2k)^{th}$,... arrivals. This creates k stochastically identical single-server queues). Since Kingman's bound is a sharper one than Brumelle's, it is important that Kingman's assertion be proved.

In this paper the case k=2 is examined in depth. We will show that

i) the limiting expected waiting time for the FIFS system is less than or equal to that for the cyclic system, however,

[1] Since the single-server queue thus generated does not have independent input, Brumelle had to extend Marshall's results.

ii) there exist realizations where the average wait is larger for the FIFS system, and

iii) there exist service time and interarrival time distributions for which the waiting times are not stochastically greater for the cyclic system.

II. The Two Systems - Definitions and Elementary Notions

We consider a system of two parallel servers each working at unit rate. The i^{th} arriving customer (denoted by $\textcircled{i}$) has service time S_i and interarrival time T_i (taken to be the time between the arrivals of $\textcircled{i}$ and $\textcircled{i+1}$). S_i and T_i are random variables. We assume that the first customer finds the system empty. Initially, no assumptions are made about the probability law governing the process $\{(S_i,T_i),\ i=1,2,\ldots\}$. A given set of numbers $\{(s_i,t_i)\ :\ s_i \geq 0, t_i \geq 0; i=1,2,\ldots\}$ is called a realization of the process if $S_i=s_i, T_i=t_i$ for all $i \geq 1$.

Under the FIFS system (hereafter denoted system I) $\textcircled{i}$ is immediately served if he finds one of the servers idle; otherwise he waits in queue an amount of time W_i until the first time a server becomes free after $\textcircled{i-1}$ starts service. Although this description suggests a single queue of waiting customers, it is more convenient, as in [2] , to consider each server to have his own queue, such that a given customer immediately joins the queue of that server who eventually serves him. If we denote by Q_i the hypothetical waiting time of $\textcircled{i}$ if he were to join the other queue, then the pair of numbers of (W_i,Q_i) denotes the remaining work (at the instant of $\textcircled{i}$'s arrival) of the two respective servers if i-1 were to be the last customer granted service. These definitions yield the following recursion relationships:

$$\begin{aligned} W_{i+1} &= \min\left\{ \left[W_i+S_i-T_i\right]^+, \left[Q_i-T_i\right]^+ \right\} \\ Q_{i+1} &= \max\left\{ \left[W_i+S_i-T_i\right]^+, \left[Q_i-T_i\right]^+ \right\} \end{aligned} \qquad (1)$$

where the $[\ ]^+$ operator denotes the diode function:

$$[x]^+ = \begin{cases} x \text{ if } x \geq 0 \\ o \text{ if } x < 0. \end{cases}$$

A useful schematic aid in depicting the process is shown in Fig. 1, with the bars being "fed" sporadically on the right by amounts S_i and being "eaten away" from the left at a uniform rate.

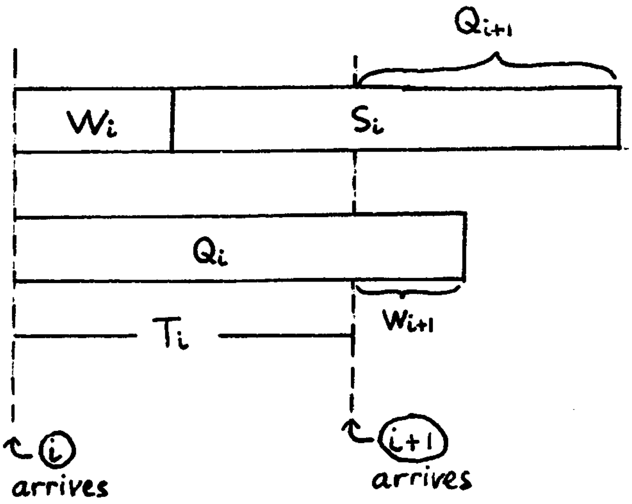

Fig. 1. Geometric Depiction of Work Loads.

In the cyclic system (hereafter denoted system II) customer $\textcircled{i}$ goes to the first server if i is odd, to the second if i is even. We define $\widetilde{W}_i$ to be the current visible work of the server that $\textcircled{i}$ goes to at the time of his arrival, $\widetilde{Q}_i$ the work of the other server, and obtain the following recursion relationships:

$$\begin{aligned} \widetilde{W}_{i+1} &= \left[\widetilde{Q}_i - T_i\right]^+ \\ \widetilde{Q}_{i+1} &= \left[\widetilde{W}_i + S_i - T_i\right]^+ . \end{aligned} \tag{2}$$

We note here briefly that system II is, in principle, simpler than system I, since it can be reduced to a one-dimensional process by considering two-step recursions:

$$\widetilde{W}_{i+2} = \left[\widetilde{W}_i + S_i - T_i - T_{i+1}\right]^+ . \tag{3}$$

It is useful to define two more quantities:

Total Work: $L_i = W_i + Q_i$

Unevenness: $U_i = Q_i - W_i$

(and, similarly, $\widetilde{L}_i$ and $\widetilde{U}_i$ in system II).

In comparing system I with system II it is at first instructive to assign them the same realization of $\left\{(S_i, T_i),\ i=1,2,\ldots\right\}$ and calculate the corresponding set of numbers

(W_i, Q_i) and $(\tilde{W}_i, \tilde{Q}_i)$ for every i. For every realization we have $W_1 = \tilde{W}_1 = W_2 = \tilde{W}_2 = 0$ and the two systems run identically until some arrival $\textcircled{i}$ $(i \geq 3)$ sees $(W_i = a, Q_i = b)$ in system I and $\tilde{W}_i = b, \tilde{Q}_i = a)$ in system II for some a,b where $a < b$. At this point the two systems part company with system I gaining an immediate advantage in waiting time but at the cost of an "inferior" state for the next arrival. A "good" state, intuitively, is one with a high amount of unevenness for a given amount of total work. However, very high unevenness may occasionally cause a server to become prematurely idle, which is "bad" since then the total work is not being reduced at maximum rate.

Intuitively, it is clear that system II will tend to have more imbalance in the two queue lengths than system I and hence be more susceptible to partial server idleness. This implies more total work and therefore longer waiting times. However, it is difficult to quantify these statements for a formal proof. In fact, the proof presented in section IV will not be along this line of reasoning.

III. The Two Systems - Some Partial Results

A. Comparison of Realizations

In this section we view the S_i and T_i as given numbers and ignore their probability laws. No restrictions are placed on them other than their non-negativity. We will say that an n-block violation occurs starting at $\textcircled{i}$ if

$$\sum_{j=i}^{i+n-1} W_j > \sum_{j=i}^{i+n-1} \tilde{W}_j$$

and ask whether it is possible for an n-block violation to occur starting at $\textcircled{1}$ and, if so, what is the minimum such n. These questions are answered as follows (the groundwork and proofs are presented in Appendix A):

Lemma 1: Suppose arrival $\textcircled{i}$ finds the state to be (a,b) in system I and (b,a) in system II with $a < b$. Then an n-block violation starting at $\textcircled{i}$ is impossible for n=1,2,or 3.

Since the condition of Lemma 1 can be fulfilled at the very earliest at i=3, it is clear that, starting at $\textcircled{1}$, 5-block violations cannot occur. However, 6-block violations are possible:

Theorem 1: The minimum n necessary for an n-block violation, starting at $\textcircled{1}$, is 6.

Let us for a moment consider the service and inter-arrival times to be independent random variables, so that the instances where an arrival finds both system I and II empty are renewal points. With a very light traffic intensity ($ES/2ET \ll 1$) the renewal periods would encompass only a few arrivals, which, by Theorem 1, would tend to favour system I.

We will say that a system becomes <u>partially idle</u> if, when one of the servers finishes the work of a customer, no new customer is in the queue to immediately take his place.

A necessary condition for a 6-block violation is that system I becomes partially idle before ⑤ arrives. One may ask whether an n-block violation ($n \geq 7$) is possible in which system I never becomes idle. The answer is no:

<u>Theorem 2</u>: If system I does not become partially idle at any time after the arrival of ② , then an n-block violation, starting at ① , cannot occur for any n.

Thus system I is certainly no worse than system II in a fully congested system (no partial idleness).

Suppose we have an n-block violation for a given set of numbers $\{(S_i, T_i),\ i=1,2,\ldots,n\}$. Let us take the n ordered numbers $S_1, S_2, \ldots, S_n$, permute them in some way and compute the new waiting times. For a permutation σ in Λ_n, the group of permutations of $\{1,..,n\}$, let $W_i^\sigma, \tilde{W}_i^\sigma$ be the corresponding waiting times obtained as above. Consider the conjecture that

$$\sum_{\sigma \in \Lambda_n} \sum_{i=1}^{n} W_i^\sigma \leq \sum_{\sigma \in \Lambda_n} \sum_{i=1}^{n} \tilde{W}_i^\sigma \quad .$$

If this conjecture was true for all n and any set of S_i, it would follow that the expected wait in system I is less than or equal to that in system II provided that the S_i are independent and identically distributed. Unfortunately, the author has been able to prove this conjecture only for the case of n=6.

B. Comparison of Distribution Functions for Waiting Times and Total Work

In this section, we make the usual assumptions about the $\{(S_i,T_i), i=1,2,...\}$ process, namely that the service times are i.i.d. with distribution function H(s) and independent of the interarrival times which are also i.i.d. with distribution function G(t). We define the following two-variable distribution functions:

$$F_i(x_1,x_2) = P\left[W_i \leq x_1, Q_i \leq x_2\right]$$

$$\tilde{F}_i(x_1,x_2) = P\left[\tilde{W}_i \leq x_1, \tilde{Q}_i \leq x_2\right] .$$

The following recursion relationships can be derived via a method given in [4] (see appendix B):

Let $U(x_1,x_2)$ be the two-dimensional unit-step-function:

$$U(x_1,x_2) = \begin{cases} 1 \text{ if } x_1 \geq 0 \text{ and } x_2 \geq 0 \\ 0 \text{ otherwise} \end{cases}$$

Then $F_1(x_1,x_2) = \tilde{F}_1(x_1,x_2) = U(x_1,x_2)$

and, for $i \geq 1$,

$$\tilde{F}_{i+1}(x_1,x_2) = U(x_1,x_2)\iint \tilde{F}_i(x_2-s+t,x_1+t)\,dH(s)\,dG(t) \qquad (4)$$

$$F_{i+1}(x_1,x_2) = \begin{cases} U(x_1,x_2)\iint F_i(x_2-s+t,x_2+t)\,dH(s)\,dG(t) \text{ if } x_1 \geq x_2 \\ U(x_1,x_2)\iint A_i(x_1,x_2,s,t)\,dH(s)\,dG(t) \text{ if } x_1 < x_2 \end{cases} \qquad (5)$$

where

$$A_i(x_1,x_2,s,t) = F_i(x_1-s+t,x_2+t)+F_i(x_2-s+t,x_1+t)-F_i(x_1-s+t,x_1+t) .$$

It is shown in [4] that $F_i(x_1,x_2) \geq F_{i+1}(x_1,x_2)$ for every (x_1,x_2) and every i, and that

$$F(x_1,x_2) = \lim_{i\to\infty} F_i(x_1,x_2)$$

is the steady-state distribution function provided that $ES < 2ET$.

A similar statement may be made for system II.

We consider the following series of conjectures (an omitted subscript indicates steady-state conditions):

A. $EW \leq E\tilde{W}$

B. $EW_i \leq E\tilde{W}_i$ for all i

C. $F(x,\infty) \geq \tilde{F}(x,\infty)$ for all x

D. $F_i(x,\infty) \geq \tilde{F}(x,\infty)$ for all i, all x

E. $F_i(x_1,x_2) \geq \tilde{F}_i(x_1,x_2)$ for $0 \leq x_1 \leq x_2$ and all i.

(Note that $E \Rightarrow D \Rightarrow C \Rightarrow A$ and $D \Rightarrow B \Rightarrow A$.)

When investigating these conjectures, the author at first thought it possible to prove conjecture E via the recursion relationships (4) and (5). However, it is not true for every possible pair of probability distributions H(s) and G(t). In fact, a counterexample can even be produced for conjecture C. This is shown in appendix B.

A random variable X is said to be stochastically larger than another random variable Y if, for every z, we have

$$P[X > z] \geq P[Y > z].$$

The fact that conjecture C does not hold for every H(s) and G(t) means that $\tilde{W}$ is not necessarily stochastically larger than W.

The author does believe that $\tilde{L}$ is stochastically larger than L. However, as indicated in appendix B, the recursion relationships involving $P\left[U_i \leq x_1, L_i \leq x_2\right]$ are not nearly as simple as those for W_i and Q_i. Moreover, the non-monotonic behavior of the unevenness quantity, alluded to in Section II and further illustrated in appendix B, makes it difficult to generate a suitable induction statement involving the U_i. Consequently, the question whether $\tilde{L}$ is stochastically larger than L is left as an unsolved problem.

IV. Proof of $EW \leq E\tilde{W}$ via Value Functions

As in section II, we consider each server to have his own queue, but assume at first some general but fixed policy of assigning arrivals to servers. We also drop the assumption of an empty system when $\textcircled{1}$ arrives. Observing this stochastic process only at the arrival epochs, we may extract the following discrete-parameter, two-variable continuous state stochastic process:

$$\left\{ Z_i=(X_i,Y_i),\ i=1,2,\ldots \;\middle|\; Z_1=(x,y) \right\}$$

where Y_i = work that $\textcircled{i}$ sees in front of that server to whom $\textcircled{i-1}$ was assigned,

X_i = work that $\textcircled{i}$ sees in front of the other server.[1]

Under a stationary policy the decision to assign $\textcircled{i}$ to a specific server depends only on the state Z_i of the process and not on the previous history. For example, both system I (FIFS) and system II (cyclic) constitute stationary policies.

We note that if, in addition to a stationary policy, we also have i.i.d. service times and i.i.d. interarrival times, then $\{Z_i,\ i=1,2,\ldots\}$ is a Markov Process. If we leave the policy open to choice and specify some cost structure, we have a Markov Decision Process. However, in the following exposition we will not assume that the

[1] This is a slight change in order from the state definition of Section III A.

interarrival times are i.i.d., and consequently the resulting process may not be Markov. Nevertheless, we will borrow from Markov Decision Process theory the concept of immediate cost and state value functions.

We restrict our attention now ~~to now~~ to two policies: FIFS and cyclic assignment. We take as immediate cost the actual waiting time and consider the criterion of total expected cost over a finite horizon. The following conditions on the $\{(S_i,T_i),i=1,2,..\}$ process will be sufficient to prove system I to be better than system II:

C_1) $\{T_i\}$ is any arbitrary sequence of non-negative random variables.

C_2) $\{S_i\}$ is an i.i.d. sequence of non-negative random variables independent of $\{T_i\}$..

For convenience, we denote by $\mathbb{R}_+^n$ the portion of Euclidean n-space in which every coordinate is non-negative. A given realization of the first n interarrival times $\underline{t}=(t_1,t_2,...,t_n)$ and the first n service times $\underline{s}=(s_1,s_2,...,s_n)$ can thus be viewed as points in $\mathbb{R}_+^n$ and the sequence of random variables $\underline{T}=(T_1,T_2,...,T_n)$ and $\underline{S}=(S_1,S_2,...,S_n)$ as random vectors in $\mathbb{R}_+^n$.

We define the following value functions for $k \leq n$:

$$V_k(x,y,\underline{t},\underline{s}) = \sum_{i=1}^{k} (W_i \mid Z_1=(x,y),\underline{T}=\underline{t},\underline{S}=\underline{s}) \tag{6}$$

$$V_k(x,y,\underline{t}) = E_{\underline{S}}\, V_k(x,y,\underline{t},\underline{S}) \tag{7}$$

$$V_k(x,y) = E_{\underline{T}}\, V_k(x,y,\underline{T}) \tag{8}$$

where W_i is the waiting time of ⓘ in system I. Corresponding definitions for system II are made using the "$\sim$" symbol. Referring back to Section III A, we see that $V_n(x,y,\underline{t},\underline{s})$ is an n-block sum of waiting times for a given realization starting with the state (x,y). We note that $V_n(x,y,\underline{t},\underline{s})=V_n(y,x,\underline{t},\underline{s})$ and also that by Lemma 1

$$V_n(x,y,\underline{t},\underline{s}) \leq \tilde{V}_n(x,y,\underline{t},\underline{s}) \quad \text{for } n=1,2, \text{ or } 3. \tag{9}$$

We will prove that

$$V_n(x,y,\underline{t}) \leq \tilde{V}_n(x,y,\underline{t}) \tag{10}$$

for any $(x,y) \in \mathbb{R}_+^2$, any $\underline{t} \in \mathbb{R}_+^n$, and any positive integer n. It is clear that if (10) holds then so does

$$V_n(x,y) \leq \tilde{V}_n(x,y). \tag{11}$$

If, in addition,

i) $\{T_i\}$ is an i.i.d. sequence, and

ii) $ES < 2ET$,

then the steady state waiting times EW and $E\tilde{W}$ exist and are given by

$$EW = \lim_{n\to\infty} \frac{1}{n} V_n(x,y)$$

$$E\tilde{W} = \lim_{n\to\infty} \frac{1}{n} \tilde{V}_n(x,y) \qquad \text{(see, for instance, Ross [5]).}$$

Thus (10) is a stronger statement than $EW \leq E\tilde{W}$.

To prove (10), some properties of the value functions are first established via the following lemmas:

Lemma 2: $\tilde{V}_n(x,y,\underline{t},\underline{s})$ is a non-decreasing function in x and y.

This is easily proved by induction using the recursion relationship:

$$\tilde{V}_n(x,y,\underline{t},\underline{s}) = x+\tilde{V}_{n-1}\left([y-t_1]^+, [x+s_1-t_1]^+, \hat{\underline{t}},\hat{\underline{s}}\right) \qquad (12)$$

where $\hat{\underline{t}} = (t_2,t_3,\ldots,t_n) \in \mathbb{R}_+^{n-1}$

and $\hat{\underline{s}} = (s_2,s_3,\ldots,s_n) \in \mathbb{R}_+^{n-1}$.

Corollary: $\tilde{V}_n(x,y,\underline{t})$ is a non-decreasing function in x and y.

Lemma 3: $\tilde{V}_n(x_1,y_1,\underline{t},\underline{s}) + \tilde{V}_n(x_2,y_2,\underline{t},\underline{s})$

$= \tilde{V}_n(x_1,y_2,\underline{t},\underline{s}) + \tilde{V}_n(x_2,y_1,\underline{t},\underline{s})$

for any $x_1,y_1,x_2,y_2,\underline{t},\underline{s}$.

Proof: This can be seen intuitively by considering two cyclic queuing systems with different initial work loads but identical future arrivals and exchanging their second servers. Formally, we proceed by induction:

n=1: Each side is equal to x_1+x_2.

We assume the statement for i=1,2,...,n-1 and use relationship (12):

$$\begin{aligned}
&\tilde{V}_n(x_1,y_1,\underline{t},\underline{s}) + \tilde{V}_n(x_2,y_2,\underline{t},\underline{s})\\
&= x_1 + \tilde{V}_{n-1}([y_1-t_1]^+, [x_1+s_1-t_1]^+,\hat{\underline{t}},\hat{\underline{s}})\\
&+ x_2 + \tilde{V}_{n-1}([y_2-t_1]^+, [x_2+s_1-t_1]^+,\hat{\underline{t}},\hat{\underline{s}})\\
&= x_2 + \tilde{V}_{n-1}([y_1-t_1]^+, [x_2+s_1-t_1]^+,\hat{\underline{t}},\hat{\underline{s}})\\
&+ x_1 + \tilde{V}_{n-1}([y_2-t_1]^+, [x_1+s_1-t_1]^+,\hat{\underline{t}},\hat{\underline{s}})\\
&= \tilde{V}_n(x_2,y_1,\underline{t},\underline{s}) + \tilde{V}_n(x_1,y_2,\underline{t},\underline{s}) \ .
\end{aligned}$$

Corollary: $\tilde{V}_n(x_1,y_1,\underline{t}) + \tilde{V}_n(x_2,y_2,\underline{t}) = \tilde{V}_n(x_1,y_2,\underline{t}) + \tilde{V}_n(x_2,y_1,\underline{t})$

Lemma 4: Let (a_1,a_2) and (b_1,b_2) be states given by

$$\begin{aligned}
a_1 &= [x+s-t_1-t_2]^+\\
a_2 &= [[y-t_1]^+ +s-t_2]^+\\
b_1 &= [\ y+s-t_1-t_2\ ]^+\\
b_2 &= [[x-t_1]^+ +s-t_2]^+
\end{aligned}$$

where $0 \leq x < y$, $s \geq 0$, $t_1 \geq 0$, $t_2 \geq 0$.

Then $a_1 \leq a_2$ and either

A: $a_1 \leq b_2$ and $a_2 = b_1$

or B: $a_1 \leq b_1$ and $a_2 = b_2$.

Proof: 1) $a_2 \geq \left[y - t_1 + s - t_2\right]^+ \geq a_1$ (using PLUS 1 and PLUS 3 from appendix A)

2) if $t_1 \leq y$ then

$a_2 = b_1$ and $b_2 \geq a_1$ (by PLUS 1, PLUS 3)

which is case A.

If $t_1 > y$ then

$a_2 = b_2$ and $a_1 \leq b_1$ (by PLUS 3)

which is case B.

The following lemma will be essential in the final proof of (10):

Lemma 5: Let conditions C_1 and C_2 hold for the arrival and service processes and let x,y be numbers satisfying $0 \leq x < y$. Then

$$\tilde{V}_n(x,y,\underline{t}) \leq \tilde{V}_n(y,x,\underline{t})$$

for any $\underline{t} \in \mathbb{R}_+^n$ and any positive integer n.

Proof: By induction.

For n=1, $\tilde{V}_1(x,y,\underline{t}) = x < y = \tilde{V}_1(y,x,\underline{t})$.

For n=2, $\tilde{V}_2(x,y,\underline{t}) = x + [y-t_1]^+$

$$\tilde{V}_2(y,x,\underline{t}) = y + [x-t_1]^+.$$

The desired inequality follows from PLUS 4. Now assume induction hypothesis for i=1,2,...,n-1. If the present state is (x,y), then the next two waiting times will be x and $[y-T_1]^+$, and the future state (two steps from now) will be

$$([x+S_1-T_1-T_2]^+, [[y-T_1]^+ + S_2-T_2]^+).$$

Hence,

$$\tilde{V}_n(x,y,\underline{t}) = x + [y-t_1]^+ + \underset{S_1,S_2}{E}\tilde{V}_{n-2}([x+S_1-t_1-t_2],[[y-t_1]^+ + S_2-t_2]^+,\hat{\underline{t}}) \tag{13}$$

where $\hat{\underline{t}} = (t_3,\ldots,t_n) \in \mathbb{R}_+^{n-2}$.

Denoting by $A(x,y,\underline{t})$ the last term of (13), we have

$$\tilde{V}_n(x,y,\underline{t}) = x + [y-t_1]^+ + A(x,y,\underline{t})$$

$$\tilde{V}_n(y,x,\underline{t}) = y + [x-t_1]^+ + A(y,x,\underline{t})$$

We show that $A(x,y,\underline{t}) \le A(y,x,\underline{t})$.

By definition,

$$A(x,y,\underline{t}) = \iint_{[(s_1,s_2):s_1 \ge 0, s_2 \ge 0]} \tilde{V}_{n-2}([x+s_1-t_1-t_2]^+, [[y-t_1]^+ + s_2-t_2]^+, \hat{\underline{t}})\, dH(s_1)\, dH(s_2).$$

Dividing the region of integration, we have

$$A(x,y,\underline{t}) = \iint\limits_{[(s_1,s_2):s_1=s_2\geq 0]} \tilde{V}_{n-2}([x+s_2-t_1-t_2]^+, [[y-t_1]^+ +s_2-t_2]^+, \hat{\hat{\underline{t}}})\, dH(s_1)\, dH(s_2)$$
$$+ \iint\limits_{[(s_1,s_2):0\leq s_1< s_2]} \Big[\tilde{V}_{n-2}([x+s_1-t_1-t_2]^+, [[y-t_1]^+ +s_2-t_2]^+, \hat{\hat{\underline{t}}})$$
$$+\tilde{V}_{n-2}([x+s_2-t_1-t_2]^+, [[y-t_1]^+ +s_1-t_2]^+, \hat{\hat{\underline{t}}})\Big]\, dH(s_1)\, dH(s_2).$$

Finally, employing the corollary to Lemma 3:

$$A(x,y,\underline{t}) = \iint\limits_{[(s_1,s_2):s_1=s_2\geq 0]} \tilde{V}_{n-2}([x+s_2-t_1-t_2]^+, [[y-t_1]^+ +s_2-t_2]^+, \hat{\hat{\underline{t}}})\, dH(s_1)\, dH(s_2)$$
$$+ \iint\limits_{[(s_1,s_2):0\leq s_1< s_2]} \tilde{V}_{n-2}([x+s_1-t_1-t_2]^+, [[y-t_1]^+ +s_1-t_2]^+, \hat{\hat{\underline{t}}})\, dH(s_1)\, dH(s_2)$$
$$+ \iint\limits_{[(s_1,s_2):0\leq s_1< s_2]} \tilde{V}_{n-2}([x+s_2-t_1-t_2]^+, [[y-t_1]^+ +s_2-t_2]^+, \hat{\hat{\underline{t}}})\, dH(s_1)\, dH(s_2)$$

The corresponding integrands of $A(x,y,\underline{t})$ can be shown to be less than or equal to those of $A(y,x,\underline{t})$. For example, the respective first integrands are:

$$\tilde{V}_{n-2}([x+s_2-t_1-t_2]^+, [[y-t_1]^+ +s_2-t_2]^+, \hat{\hat{\underline{t}}}) = \tilde{V}_{n-2}(a_1, a_2, \hat{\hat{\underline{t}}})$$

and

$$\tilde{V}_{n-2}([y+s_2-t_1-t_2]^+, [[x-t_1]^+ +s_2-t_2]^+, \hat{\hat{\underline{t}}}) = \tilde{V}_{n-2}(b_1, b_2, \hat{\hat{\underline{t}}})$$

where a_1, a_2, b_1, and b_2 are as in Lemma 4 (with $s=s_2$).

For case A (in Lemma 4), we have

$$\tilde{V}_{n-2}(a_1,a_2,\hat{\hat{\underline{t}}}) \leq \tilde{V}_{n-2}(a_2,a_1,\hat{\hat{\underline{t}}}) \leq \tilde{V}_{n-2}(b_1,b_2,\hat{\hat{\underline{t}}})$$

where the first inequality follows from the induction hypothesis and the second from Lemma 2.

For case B

$$\tilde{V}_{n-2}(a_1,a_2,\hat{\hat{\underline{t}}}) \leq \tilde{V}_{n-2}(b_1,b_2,\underline{t}) \qquad \text{by Lemma 2.}$$

Applying similar arguments to the other integrands completes the proof of the lemma.

We are now ready to prove assertion (10):

<u>Theorem 3</u>: Suppose the arrival process $\{(S_i,T_i),i-1,2...\}$ obeys conditions C_1,C_2. Then for every initial state $(x,y) \in \mathbb{R}^2_+$, every $\underline{t} \in \mathbb{R}^n_+$, and every positive integer n, we have

$$V_n(x,y,\underline{t}) \leq \tilde{V}_n(x,y,\underline{t}).$$

Proof: By induction.

For n=1, $V_1(x,y,\underline{t}) = \min(x,y)$ and

$$\tilde{V}_1(x,y,\underline{t}) = x$$

If $x \leq y$, then by inductive hypothesis

$$\begin{aligned} V_n(x,y,\underline{t}) &= x + E_{S_1} V_{n-1}([y-t_1]^+, [x+S_1-t_1]^+, \hat{\underline{t}}) \\ &\le x + E_{S_1} \tilde{V}_{n-1}([y-t_1]^+, [x+S_1-t_1]^+, \hat{\underline{t}}) \\ &= \tilde{V}_n(x,y,\underline{t}) \end{aligned}$$

If $y < x$ then

$V_n(x,y,\underline{t}) = V_n(y,x,\underline{t}) \le \tilde{V}_n(y,x,\underline{t})$ as above,

but

$\tilde{V}_n(y,x,\underline{t}) \le \tilde{V}_n(x,y,\underline{t})$ by Lemma 5.

This completes the proof.

Corollary: Under the standard assumptions of queuing theory (i.e. $\{S_i\}$ is an i.i.d. sequence, independent of $\{T_i\}$, also i.i.d., $ES < 2ET$), $EW \le E\tilde{W}$.

Proof: Via standard definitions and limit theorems (see Ross [5], chapter 5)

$EW = \lim_{n\to\infty} E\left(\frac{1}{n}\sum_{i=1}^{n} W_i\right)$ independent of initial conditions.

By definition (8)

$$V_n(x,y) = E\left(\sum_{i=1}^{n} W_i \,\middle|\, Z_1=(x,y)\right)$$

The desired result follows from Theorem 3.

BIBLIOGRAPHY

1. K. T. Marshall, "Some Inequalities in Queuing", Operations Research Vol. 16, 651-665 (1968).

2. S. L. Brumelle, "Some Inequalities for Parallel-Server Queues", Operations Research Vol. 19, No. 2, 402-413 (1971).

3. J.F.C. Kingman, "Inequalities in the Theory of Queues", J. Royal Stat. Soc., Series B, Vol. 32, 102-110 (1970).

4. J. Kiefer and J. Wolfowitz, "On the Theory of Queues with Many Servers", Trans. American Math. Soc., Vol. 78, 1-18 (1955).

5. S.M. Ross, "Applied Probability Models with Optimization Applications", Chapter 5,Holden-Day (1970)

APPENDIX A: Proofs of Realization Theorems

Initial Note: It is possible to prove Theorem 1 directly by using the recursion relationships (1) and (2) and considering the cases n=3,4,5.

For n=3, we have

$$\sum_{i=1}^{3} W_i = 0 + 0 + W_3 = \min\left\{ [S_1-T_1-T_2]^+, [S_2-T_2]^+ \right\}$$

$$\le [S_1-T_1-T_2]^+$$

$$\sum_{i=1}^{3} \widetilde{W}_i = 0 + 0 + \widetilde{W}_3 = [S_1-T_1-T_2]^+$$

The case n=4 is already long and tedious. Essentially, it involves a case by case exploration of 4 different situations corresponding to the number of ways arrivals ③ and ④ join the queues in system I.

Direct proof of n=5 was not even attempted since it not only involves a doubling of the number of situations over n=4, but also the individual comparisons are more complicated. Lemma 1 avoids these problems.

It is expedient to first list some of the properties of the $[\ ]^+$ operation:

PLUS 1: $[x]^+ \geq x$

PLUS 2: If $y \geq 0$ then $[[A]^+ - y]^+ = [A-y]^+$

PLUS 3: If $A < B$ then

$$[A-x]^+ \leq [B-x]^+$$

PLUS 4: If $A < B$ and $x \geq 0$, then

$$[A]^+ + [B-x]^+ \leq [B]^+ + [A-x]^+ .$$

Proof of Lemma 1:

We have $W_i = a$, $Q_i = b$, $\tilde{W}_i = b$, $\tilde{Q}_i = a$, $a < b$.

We must show that $\sum_{j=i}^{i+n-1} W_j \leq \sum_{j=i}^{i+n-1} \tilde{W}_j$ for n=1,2,3.

n=1 $\quad W_i = a < b = \tilde{W}_i$

n=2 $\quad W_{i+1} = \min\left\{[a+S_i-T_i]^+, [b-T_i]^+\right\} \leq [b-T_i]^+$

$$\tilde{W}_{i+1} = [a-T_i]^+$$

$$\text{So } W_i + W_{i+1} \leq a + [b-T_i]^+$$

$$\tilde{W}_i + \tilde{W}_{i+1} = b + [a-T_i]^+$$

The desired inequality follows from PLUS 4.

n=3 $\quad \tilde{W}_{i+2} = [b+S_i-T_i-T_{i+1}]^+$

$W_{i+2} \leq [b-T_i-T_{i+1}]^+$ if $\textcircled{i+1}$ goes to the same server as $\textcircled{i}$ in system I.

$W_{i+2} \leq [a+S_i-T_i-T_{i+1}]^+$ if $\textcircled{i+1}$ goes to the other server.

In any case we have $W_{i+2} \leq \tilde{W}_{i+2}$ by PLUS 3. This yields the desired inequality when combined with the result for n=2.

Proof of Theorem 1:

Without loss of generality we may assume that if arrival $\textcircled{i}$ finds $W_i=Q_i$ in system I he goes to the server that $\textcircled{i-1}$ did not go to.With this convention system I and II exhibit identical behavior (including identical waiting times) until an arrival $\textcircled{i}$ sees $\tilde{W}_i > \tilde{Q}_i$ in system II. At this point we must have $W_i=\tilde{Q}_i$ and $Q_i=\tilde{W}_i$, establishing the condition of Lemma 1. The lowest i for which this can happen is i=3. So 5-block violations, starting at $\textcircled{1}$, are impossible.

To complete the proof we must show that 6-block violations can occur. We do this by exhibiting an example:

i	S_i	T_i	W_i	Q_i	$\tilde{W}_i$	$\tilde{Q}_i$
1	2	0	0	0	0	0
2	1	0	0	2	0	2
3	1	0	1	2	2	1
4	2	3	2	2	1	3
5	2	0	0	1	0	0
6	1	3	1	2	0	2
Σ			4		3	

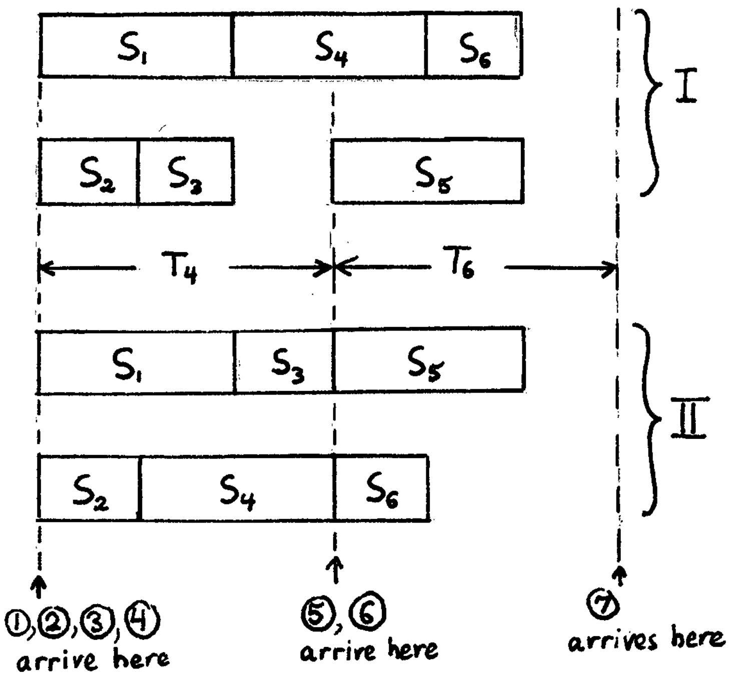

Fig. 2: Pictorial Representation of Counterexample.

Proof of Theorem 2:

We will prove a stronger result:

$W_i + W_{i+1} \leq \tilde{W}_i + \tilde{W}_{i+1}$ for any i.

To show this, we note that since system I is never partially idle (after ②), the total work in I can never be greater than that in II.

Thus we have $\tilde{L}_i = L_i + \Delta_i \quad (\Delta_i \geq 0)$.

Now, $Q_i = L_i - W_i$

and $W_{i+1} \leq \left[Q_i - T_i\right]^+ = Q_i - T_i$ (since no idleness occurs).

So $W_i + W_{i+1} \leq W_i + L_i - W_i - T_i = L_i - T_i$.

In system II, we have $\tilde{Q}_i = L_i + \Delta_i - \tilde{W}_i$

and $\tilde{W}_{i+1} = \left[\tilde{Q}_i - T_i\right]^+ \geq \tilde{Q}_i - T_i$ (by PLUS 1)

so that $\tilde{W}_i + \tilde{W}_{i+1} \geq \tilde{W}_i + L_i + \Delta_i - \tilde{W}_i - T_i = L_i + \Delta_i - T_i$,

which completes the proof.

APPENDIX B: Distribution Functions for Waiting Times and Total Work

1. Waiting Times

In Kiefer and Wolfowitz [4] recursion relationships for the k-server queue are established via the Ψ -set. In accordance with the terminology of our problem, we define Ψ-set as follows:

$$\Psi(x_1,x_2,s,t) = \Big\{(y_1,y_2): \big[W_i=y_1,Q_i=y_2,S_i=s,T_i=t\big] \Rightarrow \big[W_{i+1}\leq x_1,Q_{i+1}\leq x_2\big]\Big\}.$$

With this definition, we have

$$F_{i+1}(x_1,x_2) = \iint P\big[(W_i,Q_i)\in\Psi(x_1,x_2,s,t)\big]\ dH(s)dG(t)$$

A typical Ψ -set is shown Fig. 3.

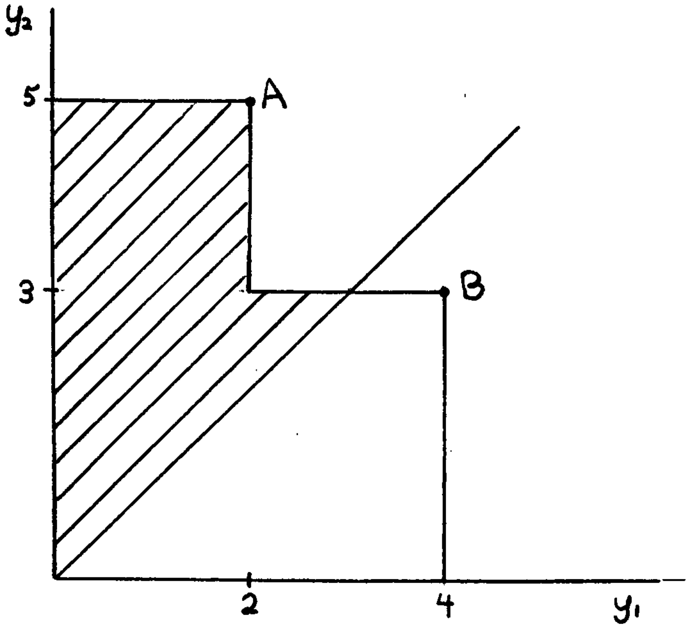

Fig. 3. Ψ -set for (W,Q)
Shaded area= $\Psi(1,3,1,2)$
$A=(x_1-s+t,x_2+t)$

$B=(x_2-s+t,x_1+t)$

The rectangles formed by the two axes and the points A and B (as shown) intersected with the region above the line $y_2=y_1$ determine the actual Ψ-set. Using the inclusion-exclusion principle and noting that there is no probability mass below the $y_2=y_1$ line, we get

$$P\left[(W_i,Q_i)\in\Psi(x_1,x_2,s,t)\right]$$

$$=F_i(x_1-s+t,x_2+t)+F_i(x_2-s+t,x_1+t)-F_i(x_1-s+t,x_1+t).$$

If $x_1\geq x_2$ then $\Psi(x_1,x_2,s,t)=\Psi(x_2,x_2,s,t)$.

This establishes equation (5).

The corresponding definition may be used for system II (replacing W_i and Q_i by $\tilde{W}_i$ and $\tilde{Q}_i$); here the resulting $\tilde{\Psi}(x_1,x_2,s,t)$ set is a complete rectangle with the upper right vertex at (x_2-s+t,x_1+t).

Hence $P\left[(W_i,Q_i)\in\tilde{\Psi}(x_1,x_2,s,t)\right]=\tilde{F}_i(x_2-s+t,x_1+t)$ thereby establishing equation (4).

2. Counterexample to conjecture C

Let H(s) and G(t) be such that

$$P\left[S_i=s\right]=\begin{cases}1/2 & s=2\\ 1/2 & s=4\\ 0 & \text{otherwise}\end{cases}$$

$$P\left[T_i = t\right] = \begin{cases} 99/100 & t=1 \\ 1/100 & t=A \text{ (large)} \\ 0 & \text{otherwise} \end{cases}$$

With this arrival process both system I and system II cannot become partially idle until for some i the event $\left[T_i=A\right]$ takes place. We pick A so large that the probability that one of the two systems will not empty out completely at that time is negligible. Thus, until $\left[T_i=A\right]$ happens for some i, the total work in system I and II grows continually and equally, with system II tending to be more unbalanced, so that it will tend to have both short and long waiting times (relative to system I).

For example

$$P\left[W_6=0 \mid T_i=1, i=1,\ldots,5\right] = 3/16 \cong F_6(0,\infty),$$

while $P\left[\tilde{W}_6=0 \mid T_i=1, i=1,..,5\right] = 1/4 \cong \tilde{F}_6(0,\infty)$.

Thus the distribution functions for the W_i and $\tilde{W}_i$ variables ($F_i(x,\infty)$ and $\tilde{F}_i(x,\infty)$) should intersect for sufficiently high values of i. Since we have a renewal process here (with the event $T_i=A$ initiating a renewal period), the intersecting behavior of $F_i(x,\infty)$ and $\tilde{F}_i(x,\infty)$ should carry over to the steady-state distribution functions $F(x,\infty)$ and $\tilde{F}(x,\infty)$.

3. Recursion Relationships Involving Work and Unevenness

With L_i and U_i defined as in section II and letting

$$K_i(x_1,x_2) = P\left[U_i \le x_1, L_i \le x_2\right],$$

we have $K_{i+1}(x_1,x_2) = \iint P\left[(U_i,L_i) \in \Psi(x_1,x_2,s,t)\right] dH(s)dG(t)$

where $\Psi(x_1,x_2,s,t) = \Big\{(y_1,y_2) : \left[U_i=y_1, L_i=y_2, S_i=s, T_i=t\right]$

$$\Rightarrow \left[U_{i+1} \le x_1, L_{i+1} \le x_2\right]\Big\}.$$

A typical Ψ-set is shown in Fig. 4.

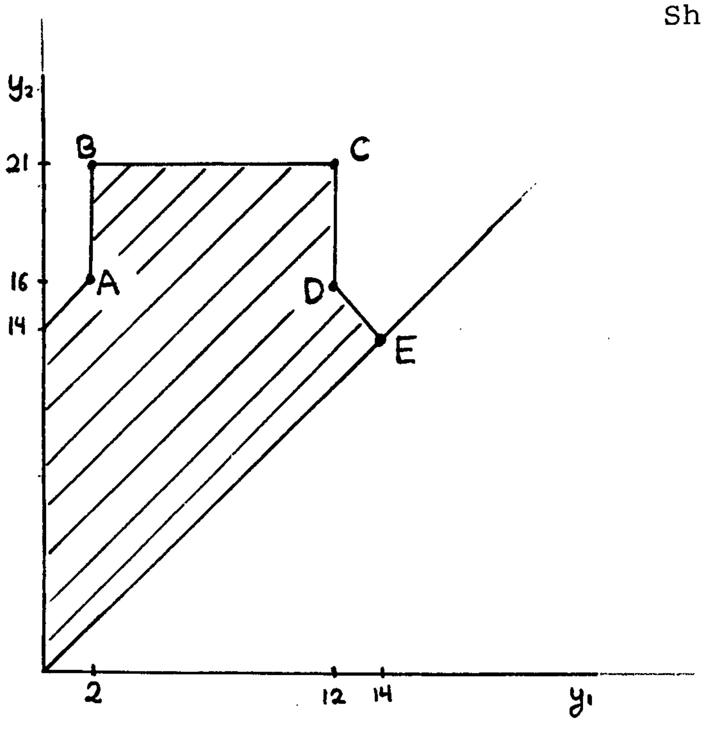

Fig. 4. Ψ-set for (U,L).

Shaded area = $\Psi(5,10,7,9)$

The vertices A-E are all functions of x_1, x_2, s, t. It is clear that the term $P\left[(U_i, L_i) \in \Psi(x_1, x_2, s, t)\right]$ cannot be expressed directly in terms of $K_i(,)$ as was the case for (W_i, Q_i). Rather such an expression would have to take the form of a double integral with very complicated limits.

A further potential problem in proving $K_i(\infty, x) \geq \tilde{K}_i(\infty, x)$ is that this statement is in itself an insufficient induction hypothesis: we would need a stronger statement such as $K_i(x_1, x_2) \geq \tilde{K}_i(x_1, x_2)$ to insure that the induction step goes through. However, the last hypothesis cannot be true since $\tilde{U}_i$ can be negative while U_i cannot be. A workable replacement is hard to find because functions involving the unevenness quantity tend to be non-monotonic. As an example consider the following function:

$$f(x,y)=E_{S,T}\left[(\tilde{L}_{i+1}-L_{i+1}) \,\middle|\, \tilde{L}_i=L_i=A, U_i=x, \tilde{U}_i=y\right]$$

This is the expected difference in the next state work given that the present work is equal but with system I and II having unevenness x and y, respectively. The non-monotonic behavior in x and y can be demonstrated even for the M/M/2 queue, with

$$H(s) = 1 - e^{-s/3}$$

$$G(t) = 1 - e^{-t/2}$$

This is shown in Fig. 5, which is a sketch of the sign of f(x,y).

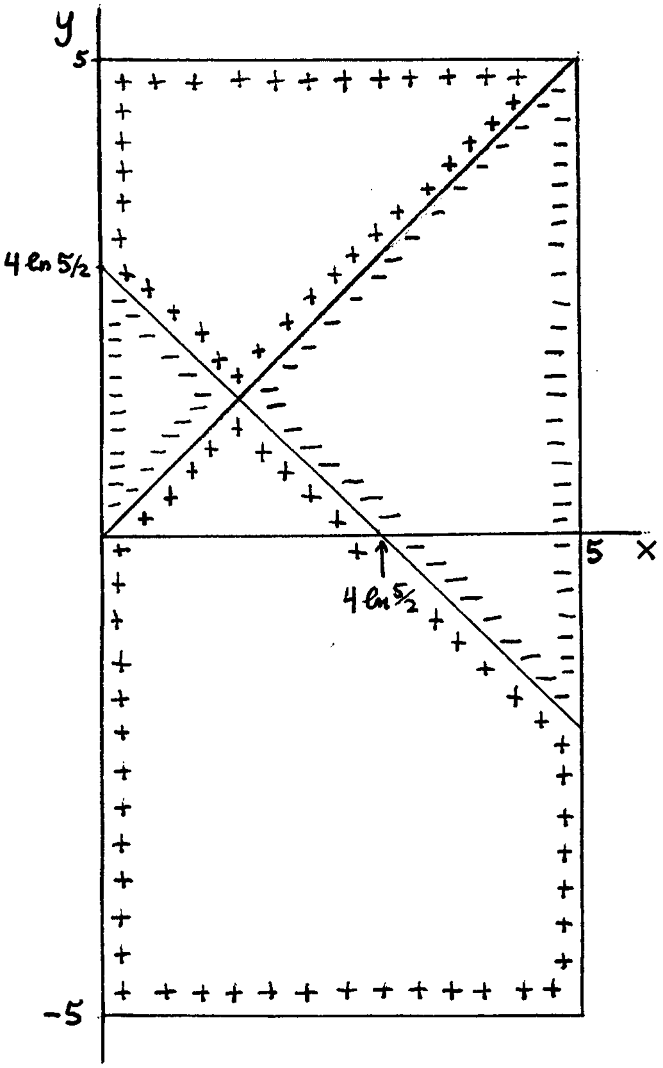

Fig. 5. An Example of the Non-Monotonic Behavior of U:

Regions where f(x,y) is positive and negative (M/M/2,A=5)